CORRELATIONS BETWEEN

VOCABULARY SIMILARITY AND INTELLIGIBILITY

Joseph E. Grimes

On the face of it, it seems reasonable that dialects whose vocabularies are similar ought to be able to understand each other rather well. Yet all too often they do not. In spite of that, decisions about language programs are sometimes made on the basis of this plausible sounding but shaky assumption.

At the low end of the scale there is a constant relationship: comprehension is always poor when vocabulary similarity is low. But that relationship does not hold up at the high end of the scale, which is where the program decisions have to be made.

The reason why high similarity is a poor predictor of high intelligibility is that there are other factors besides similarity in vocabulary that influence intelligibility. Even when vocabulary similarity is high, they can get in the way — the effect of differences in function words and affixes, syntactic and morphological rearrangements, certain kinds of regular sound shifts, and semantic shifts in both genetically derived vocabulary and loans.

Because intelligibility is so complex, the hopes raised thirty years ago by glottochronology should have faded by now; yet they have not. From one or two hundred forms, some still would see linguistic and cultural history laid out, and would decide at a glance where the paths of communication lie open. One reason the hope stays alive is that it is thought to be more work to calibrate intelligibility accurately than it is to collect a word list and compare it with other word lists, so a short cut would appear to be welcome.[1] But the energy wasted on invalid short cuts could be better employed to give us surveys that are done right the first time.

Even when intelligibility tests are given, survey reports occasionally complicate the picture by confusing intelligibility with something quite different: what amounts to bilingual behavior on the part of some of the people tested. When people learn another language, even one that is thought of as a dialect of their own language but is different enough that they cannot treat it as a simple extension of their own mother tongue, all our testing has to be done differently.

The difference comes from the fact that when a community learns a second form of speech, each person in that community does so for his or her own reasons. Some don't feel they need it, and don't learn it; some would like to, but have no opportunity; most learn it well enough for their immediate ends, but no better. This means that bilingual proficiency within a community normally varies greatly from one person to another. The sample needed to test that variation has to cover all the segments of the society, because different social sectors take differently to learning another language or dialect.

Not so with inherent intelligibility. It is an extension of ability to use the mother tongue. As a consequence, what is accessible to one member of the community is accessible to all. Its range of individual variation is fairly narrow,[2] and a smaller sample is statistically adequate for estimating it.

So in reviewing how well or how poorly intelligibility might be predicted by vocabulary similarity, we do well to remember that when bilingual comprehension is

reported as "intelligibility", we are really dealing with something whose distribution in society is quite different from that of intelligibility. In that case the degree of understanding available to those who have not gone out of their way to learn the other form of speech is lower than the figure given as if it represented uniform intelligibility.[3]

Philippines

Vocabulary similarity estimates and intelligibility test results -- subject to the cautions just given -- are available for 55 pairs of dialects in the Philippines. In these pairs the intelligibility is only weakly correlated with the vocabulary similarity.

The data are from the Tenth Edition of the *Ethnologue* (B. Grimes 1984), reproduced in Table 1 and displayed in Figures 1 and 2. They are based on field surveys made by the Philippines Branch of the Summer Institute of Linguistics. The fractions of a percentage point given in the *Ethnologue* are rounded off here to reflect better the level of accuracy that tests of the kind given yield.[4]

Seven other dialect pairs were left out because the figures reported for them in the *Ethnologue* for intelligibility are known to involve a substantial amount of bilingualism, yet they come from tests on samples that were too small to be valid for the variation that goes with bilingualism. The languages involved are not even in the same linguistic subgroupings; on purely comparative grounds it would be strange if they understood each other inherently. Atta of Pamplona tested on Ilocano had a similarity of 83% and what was called "intelligibility" of 85%; Itawit on Ilocano, 53% and 68%; Kasiguranin on Tagalog, 52% and 92%; Agusan Manobo on Cebuano, 81% and 88%; Obo Manobo on Cebuano of Nasuli, 35% and 78%; Central Tagbanwa on Cuyonon, 48% and 61%; and Central Tagbanwa on Tagalog, 40% and 54%. Ilocano, Tagalog, and Cebuano are used over wide regions of the Philippines, and Pilipino, based mainly on Tagalog, is taught in the schools.

It is likely that unrecognized bilingualism is a factor in some of the other samples reported here as well. If it is, the degree of understanding of the part of the population that has not learned the other language very well is sure to be lower than the averages suggest.

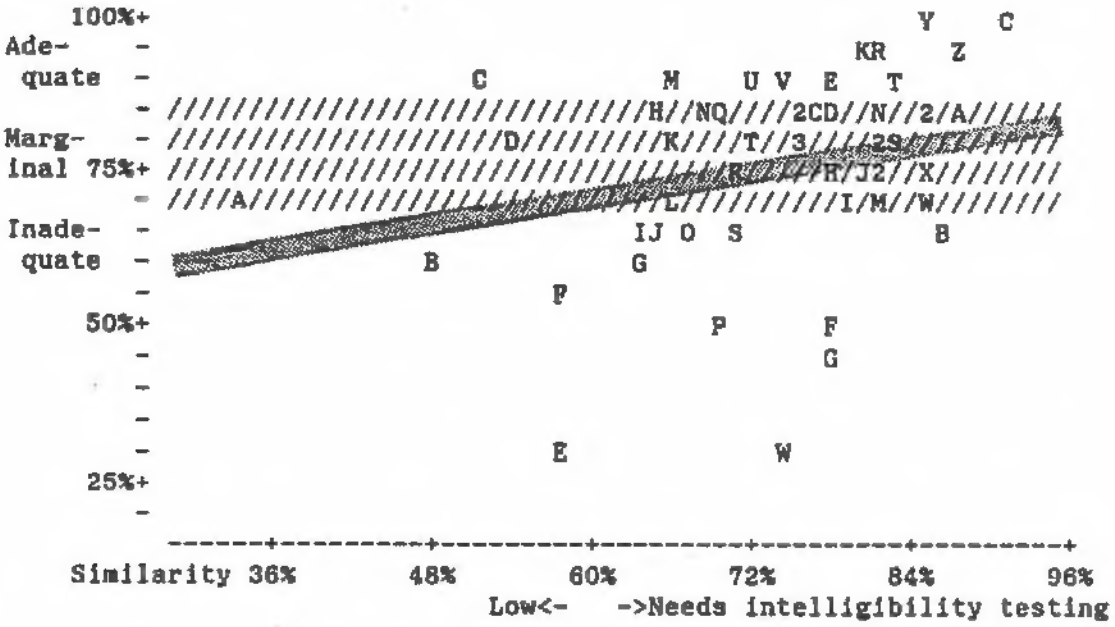
There is also a question about how the vocabulary similarity figures were arrived at. In some cases -- I do not know which -- the figures probably represent the proven cognates that remain between two word lists after borrowing and internal analogies have haphazardly upset the smooth progress of sound change. Such cognate judgments would be based on the extensive studies of phonological comparison that have been made in parts of the Philippines. In most cases, however, I take the figures to represent proportions derived from impressionistic counts of words that appear to be phonetically similar, without any way to distinguish those that come from a single parent form at an earlier historical stage via demonstrable sequences of sound changes on the one hand, and loans and analogical formations on the other.[5]

Figure 1 shows how vocabulary similarity figures relate to intelligibility measurements in the Philippines. The fifty-five dialects in Table 1, from which Figure 1 is derived, are arranged from low to high similarity. They are identified with the letters A, B, C, and so forth for the first 26 on the list, corresponding to the left part of the figure, again as A, B, C,... for the next 26 going toward the right, and A, B, C for the last three on the extreme right. Where two or three dialects fall on the

same point, a digit is given instead of a letter to show how many dialects are there. The graphs are unretouched output from the Minitab computer program, with clarifying annotations added.

Figure 1. Vocabulary similarity and intelligibility in 55 dialect pairs in the Philippines.

A,B,... identify rows in Table 1, and are repeated after the 26th and 52nd dialects on the list.
Numbers are used where dialects appear on the same spot.
The zone of marginal intelligibility is shown by ////////////////.
The regression line for the relation is shown by ▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒▒.



Normally (Simons 1979) vocabulary similarity percentages of 60% and below go consistently with intelligibility measured at 67% and below on simple narrative material. That level, for practical purposes, is inadequate for all but the simplest communication. Intelligibility seems to have to be above 85%, as measured on narrative, before much complex and personally revealing communication is likely to take place; Casad's discussion of Kirk's validation tests for Mazatec (Casad 1974.83-86) points to a 90% threshold for being able to extrapolate from a test on narrative to more complex kinds of communication.

It is clear from Figure 1 that similarity figures above Simons's 60% similarity line go with a wide range of intelligibility: from 31% (W) in the lower right to 98% (C) in the upper right. They are evenly balanced in that ten indicate adequate intelligibility (90% and up) and ten indicate inadequate intelligibility (under 70%). What relation there is between vocabulary similarity (s) and intelligibility (i) is indicated by the slanting shaded line, whose equation is

$$i = 0.465s + 41.8.$$

But the scatter of the dialects away from the line, measured by a correlation of 0.34, shows that there is only a weak relation between the two scales.

The actual dialect pairs for the Philippines that are displayed in Figure 1 are given in Table 1, with the vocabulary similarity and intelligibility figures for each. They begin with the lowest similarity figures in order to make it easy to visualize how the corresponding intelligibility scores vary.

Table 1. Philippine dialect pairs, with vocabulary similarity and intelligibility data from the Ethnologue

	VOCAB	INTELL	DIALECT TESTED on DATA FROM REFERENCE DIALECT
A	34	89	Bagobo on Tagabawa Manobo
B	48	61	Central Tagbanwa on Cuyonon
C	52	90	Agutaynon on Calamian Tagbanwa
D	54	78	Cuyonon on Tagbanwa
E	57	29	Central Tagbanwa on Lamane
F	57	56	Central Tagbanwa on Calamian Tagbanwa
G	63	60	Obo on Tagabawa
H	65	83	Madukayang on Balangao
I	65	66	Tagbanwa on Quezon Palawano
J	66	83	Yaga on Central Cagayan Agta
K	66	82	Mt. Iriga Agta on Mt. Iraya Agta
L	66	72	Mt. Iriga Agta on Central Bicolano
M	66	92	Kamayo on Surigaonon
N	68	83	Madukayang on Limos
O	68	66	Ambala on Botolan Sambal
P	69	52	Pamplona Atta on Itawit
Q	70	87	Butuanon on Kamayo
R	71	73	Aklanon on Hiligaynon
S	72	64	Ibatan on Itbayatan Ivatan
T	72	78	Batad on Kiangan Ifugao
U	72	91	Piso on Kagan Kalagan
V	74	92	Mansaka on Kagan Kalagan
W	74	31	Ibatan on Basco Ivatan
X	75	82	Kasiguranin on Paranan
Y	75	85	Sibuco-Vitali on Balangingi Sama
Z	76	86	Mt. Iriga Agta on Iriga Bicolano
A	76	78	Karao on Ibaloi
B	76	81	Rajah Kabungsuan Manobo on San Miguel Calatugan Agusan
C	77	84	Lutungan on Balangingi Sama
D	78	87	Ayangan on Batad Ifugao
E	78	88	Hapao on Kiangan Ifugao
F	78	48	Tanudan on Butbut
G	78	47	SW Palawano on Central Palawano
H	78	76	SW Palawano on Quezon Palawano

I 79	70	Butbut on Guinaang Kalinga
J 80	77	Amganad on Kiangan Ifugao
K 80	94	Calamian Tagbanwa on Baras
L 81	81	Guinaang on Balbalasang
M 81	70	Guinaang on Limos
N 81	86	Madukayang on Mangali
O 82	74	Bangad on Butbut
P 82	81	Rajah Kabungsuan Manobo on Dibabawon
Q 82	76	Brooke's Point Palawano on Quezon Palawano
R 82	96	Brooke's Point Palawano on Central Palawano
S 83	81	Burnay on Amganad Ifugao
T 83	88	Brooke's Point Palawano on Southern Palawano
U 85	87	Mayoyaw on Batad Ifugao
V 85	83	Agusan on Dibabawon
W 85	69	Brooke's Point Palawano on SW Palawano
X 85	76	SW Palawano on Brooke's Point Palawano
Y 85	98	Palanan Dumagat on Paranan
Z 87	94	Casiguran Dumagat on Paranan
A 87	85	Hungduan on Kiangan Ifugao
B 87	63	Sindanga on Tuboy-Salog
C 91	98	Pamplona Atta on Northern Ibanag

Correlations

Table 2 and the condensed counterparts of it that are given in Table 3 for other data pinpoint a few areas for which there actually is a high correlation between vocabulary similarity and intelligibility over a part of the range. As is plain from Table 2, occasional areas of high correlation show up in very limited combinations of similarity and intelligibility, and comparing it with the tables given later shows that it cannot be predicted generally from one language area to another.

In Table 2 the data are the Spearman *r* correlations for all dialect pairs whose vocabulary similarity is greater than or equal to the threshold percentage given at the bottom of the column, and whose intelligibility test results at the same time are greater than or equal to the threshold given at the left of the row. The correlations are given to two decimal places in the (a) part of Table 2, and in a condensed form of one digit with the decimal point removed in the (b) part, which is also the format of Table 3.

The asterisks (*) represent those parts of the table for which there are fewer than 5 pairs available. From fewer pairs it is impossible to calculate a meaningful correlation.

The table focuses on correlations in the area of interest, from 60% and up for vocabulary similarity and from 70% and up for intelligibility. No figures are given for 95% and up; at that level there are not enough instances in any of the data sets to produce a correlation. An additional row and column have been added in order to include the full range of levels in the table.

The lower left corner cell (0,0) of Table 2 gives the correlation for all the data. In effect it measures the entire scatter away from the line of regression that Figure 1 shows: 0.34.

If we look only at that part of the data where Simons expects to find a useful level of intelligibility, 60% lexical similarity and higher and 70% intelligibility and

higher (the 60% column with the 70% row), the scatter is relatively greater and the correlation measure is closer to zero: 0.21.

The other figures in the table show what the scatter would be if we were to cover up the lower part of Figure 1 so as to focus only on higher levels of intelligibility, and cover up the left hand part so as to focus on higher levels of vocabulary similarity, calculate the equation for another regression line like the one that crosses Figure 1 at an angle, but that covers only the points that have not been covered up, then calculate how much those data points deviate from what the new equation predicts they should be. For example, if we look only at those pairs of dialects whose similarity measures are 65% or more (the 65% column in Table 2) and whose intelligibility measures are 80% or more (the 80% row in Table 2), the correlation measure that indicates how well the equation predicts the facts is 0.31.

Perfect correlation would have been registered if all the data in a plot like Figure 1 had fallen on a straight line instead of scattering out all over the plot as they have. In Table 2 perfect correlation would be expressed by a figure of 1.00.[6] Coefficients of correlation below 0.50 indicate considerable scatter of points off the regression line and are unimpressive in arguing for a consistent relationship between two variables; correlations below 0.30 indicate a still vaguer relationship, more scatter than the 0.34 that Figure 1 demonstrates. The computer program that calculated the 1,953 correlations that went into Tables 2 and 3 is reproduced as an appendix.[7]

Table 2. Philippine dialect pairs,
correlations of vocabulary similarity and intelligibility
from the Ethnologue

All dialect pairs whose similarity is equal to or higher than the figure at the bottom, and whose intelligibility is equal to or higher than the figure at the left, are correlated:

(a) to two decimal places, (b) condensed

(a)

PHILIPPINES.DATA

Intelligibility>=I%	0%	60%	65%	70%	75%	80%	85%	90%
90%	0.85	0.83	0.83	0.85	0.52	0.52	*	*
85%	0.22	0.30	0.30	0.41	0.54	0.24	0.37	*
80%	0.19	0.31	0.31	0.28	0.49	0.37	0.48	*
75%	0.18	0.18	0.18	0.24	0.40	0.40	0.53	*
70%	0.14	0.21	0.21	0.24	0.36	0.46	0.53	*
0%	0.34	0.29	0.25	0.23	0.23	0.19	0.34	*

Similarity>=S% for 55 pairs

* N<5 too few to correlate

(b)

PHILIPPINES.DATA (55)

Intelligibility>=I%	0%	60%	65%	70%	75%	80%	85%	90%
90%	8	8	8	8	5	5	*	*
85%	2	2	2	4	5	2	3	*
80%	1	3	3	2	4	3	4	*
75%	1	1	1	2	4	4	5	*
70%	1	2	2	2	3	4	5	*
0%	3	2	2	2	2	1	3	*
Similarity>=S%	0	0	0	5	0	5	0	%
	0	6	6	7	7	8	8	9

0 to 9 R=.0 to .9

The overall correlation improves in places as some higher levels of vocabulary similarity and intelligibility are considered, but at best it indicates a loose relation, not one with high predictive value. There is an exception in the 90% row of the table when similarity values lower than 75% are taken into account; suddenly everything appears closely correlated. The data in question are the top three rows of Figure 1, those pairs with intelligibility 90% or higher, which can be seen to fall fairly close to a straight line. The dialects involved, those in the upper left of Figure 1, are probably showing the effects of bilingualism. No analogous localized pocket of high correlation shows up in any of the ten other sets of data investigated in this way. It therefore appears to be a local fluke that can be attributed to a few situations where bilingual behavior was not recognized; it is probably not the manifestation of any principle.

Other language areas

Similar pairings of figures are available in Simons's monograph for ten other areas of the world. As with the Philippines, in most cases it is not possible to know whether the vocabulary similarity figures are based on counts of genetically demonstrable cognates or on apparent phonetic similarity alone, nor can we be sure that the intelligibility tests were not applied to bilingual behavior by mistake. Some of the intelligibility testing was done before the internal safeguards described by Casad (1974) were developed: the Iroquois tests, for example, were the first ever given.

The data are taken from Simons, who also discussed the sources and their quality. In order to show how vocabulary similarity and intelligibility correspond or fail to correspond in general I have analyzed his data both with and without various adjustments he proposes. With one exception, the adjustment factors influence the overall picture hardly at all.

One of his adjustments is for discrepancies greater than 10% in intelligibility measurements made in two directions, village A tested on village B and village B tested on village A. Intelligibility scores are almost always asymmetric in this respect. Simons suggests that discrepancies greater than 10% are due to social factors rather than linguistic factors. Since the social factors he refers to are more or less equivalent to the bilingual learning I said sometimes takes place between closely related dialects, a threshold on the order of the 10% he suggests is one way to recognize those factors tentatively, though its magnitude needs to be validated in areas like Spanish vs. Portuguese and in Chinese dialects, where difficulties in intelligibility can be traced to specific areas of phonology. (In the data from the Philippines given in Table 2, none of the intelligibility data report tests given in two directions. It is possible that only the higher score for a pair of tests was reported.)

The adjustment Simons makes for asymmetry is to exclude the pair of dialects with the higher intelligibility score, reasoning that the lower score is less likely to reflect a bilingual learning factor. In the data marked "exclude" I follow his practice.

A major difference between Simons's correlations and mine, however, is that he includes the measure of a dialect on itself -- often called the "home town" measure -- among the data to be correlated, and I do not. Discrepancies between home town scores actually measured and the 100% we might expect are part of the information

needed in order to calibrate the test itself, but they are not part of the statement of the problem I am addressing. Even where the test results average below 100%, the effect of including the home town scores is a considerable increase in the correlations. I have therefore left all home town scores out of the calculations.

Because the earlier test designs (including the ones Casad reports) had not eliminated the sources of low home town scores,[8] it was thought necessary to adjust average scores by applying a correction factor based on the low home town score. Two kinds of corrections were applied. If it was assumed that speakers learned to take the test by the time they had finished the home town part of it, the home town score could be safely taken as not distinct from 100%, and the other scores could be left alone. Simons applies this kind of correction to his data for Biliau, Ethiopia, Mazatec, Siouan, and Trique. In Table 3 the results of adjustment for these are the same as for the basic data.

If on the other hand it was assumed that the difficulty carried over from one test into another, then the home town score was treated as equivalent to 100% and others were adjusted up in proportion to it. It is corrections of this form that Simons applied to Buang, Iroquois, Polynesia, Uganda, and Yuman. In Table 3 the results of adjustment for these are too small to notice.

Adjustments of this kind ought not to be applied willy-nilly to data from other parts of the world without independent proof that conditions sufficient to justify them hold there; often they do not. Improved criteria for test construction have for practical purposes eliminated the need for adjusting scores.

I therefore give in Table 3 three kinds of summaries of the correlation ranges in Simons's data as he gives them in his Appendix 1. The first, tagged as *DATA*, has no special adjusting factors applied. In that form the summaries are typical of the information that becomes accessible during the course of many language surveys at a stage before some of the possible adjusting factors can be estimated.

Second, for tests where other scores are adjusted proportionally to the change made in the home town score, data from Simons's "adjusted intelligibility" column are used in tables tagged as *ADJUST*.

Finally, in the tables tagged with *EXCLUDE* I exclude from the computation the scores he marks with "X" because they are more than 10% higher in intelligibility than the score going the other way. The small tables that make up Table 3 are the (b) or condensed form of the tables that were explained in connection with Table 2. Each adds in parentheses the total number of dialect pairs available when all levels are taken into account.

The negative *r* correlations in Buang, Polynesian, and Siouan indicate a reverse relation: the higher the similarity, the worse the intelligibility, within the thresholds given. Negative correlations are shown as minus signs in Table 3.

Table 3. Correlation ranges in ten language areas.
Data from Simons 1979, Appendix 1.

Tables are in the condensed (b) format of Table 1.
* indicates fewer than 5 pairs, insufficient to correlate
- indicates negative correlation
DATA unadjusted
ADJUST adjustment based on home town score
EXCLUDE exclusion of asymmetries over 10%

BILIAU.DATA (6)	BILIAU.ADJUST (6)	BILIAU.EXCLUDE (3)
Intelligibility>=I%	Intelligibility>=I%	Intelligibility>=I%
90% * * * * * * * *	90% * * * * * * * *	90% * * * * * * * *
85% 2 2 2 2 2 2 * *	85% 2 2 2 2 2 2 * *	85% * * * * * * * *
80% 3 3 3 3 3 3 * *	80% 3 3 3 3 3 3 * *	80% * * * * * * * *
75% 3 3 3 3 3 3 * *	75% 3 3 3 3 3 3 * *	75% * * * * * * * *
70% 3 3 3 3 3 3 * *	70% 3 3 3 3 3 3 * *	70% * * * * * * * *
0% 3 3 3 3 3 3 * *	0% 3 3 3 3 3 3 * *	0% * * * * * * * *
0 6 6 7 7 8 8 9	0 6 6 7 7 8 8 9	0 6 6 7 7 8 8 9
0 0 5 0 5 0 5 0%	0 0 5 0 5 0 5 0%	0 0 5 0 5 0 5 0%
Similarity>=S%	Similarity>=S%	Similarity>=S%
0 to 9 R=.0 to .9	0 to 9 R=.0 to .9	0 to 9 R=.0 to .9

BUANG.DATA (18)	BUANG.ADJUST (18)	BUANG.EXCLUDE (12)
90% * * * * * * * *	90% * * * * * * * *	90% * * * * * * * *
85% * * * * * * * *	85% * * * * * * * *	85% * * * * * * * *
80% - - - - - - * *	80% - - - - - - * *	80% * * * * * * * *
75% 4 4 4 - - - * *	75% 4 4 4 - - - * *	75% 5 5 5 * * * * *
70% 5 5 4 - - - * *	70% 5 5 4 - - - * *	70% 5 5 5 * * * * *
0% 5 5 5 3 3 0 * *	0% 5 5 5 3 3 0 * *	0% 7 7 6 7 7 4 * *
0 6 6 7 7 8 8 9	0 6 6 7 7 8 8 9	0 6 6 7 7 8 8 9
0 0 5 0 5 0 5 0%	0 0 5 0 5 0 5 0%	0 0 5 0 5 0 5 0%

ETHIOPIA.DATA (25)	ETHIOPIA.ADJUST (25)	ETHIOPIA.EXCLUDE (18)
90% * * * * * * * *	90% * * * * * * * *	90% * * * * * * * *
85% * * * * * * * *	85% * * * * * * * *	85% * * * * * * * *
80% * * * * * * * *	80% * * * * * * * *	80% * * * * * * * *
75% * * * * * * * *	75% * * * * * * * *	75% * * * * * * * *
70% * * * * * * * *	70% * * * * * * * *	70% * * * * * * * *
0% 6 9 * * * * * *	0% 6 9 * * * * * *	0% 5 9 * * * * * *
0 6 6 7 7 8 8 9	0 6 6 7 7 8 8 9	0 6 6 7 7 8 8 9
0 0 5 0 5 0 5 0%	0 0 5 0 5 0 5 0%	0 0 5 0 5 0 5 0%

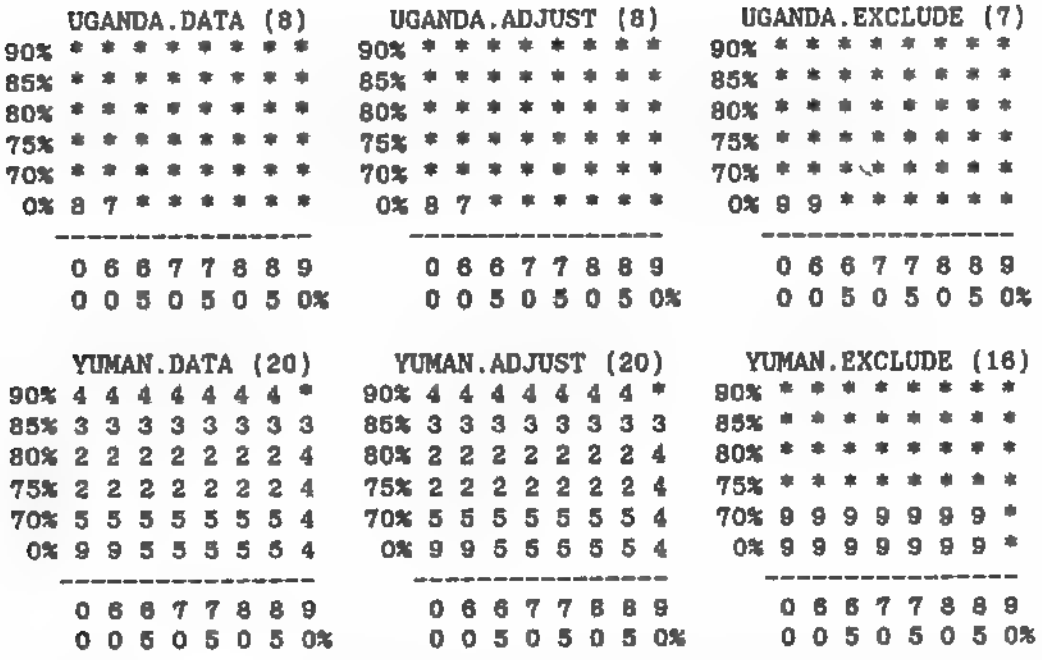
```
  IROQUOIS.DATA (10)        IROQUOIS.ADJUST (10)      IROQUOIS.EXCLUDE (8)
90% * * * * * * * *       90% * * * * * * * *       90% * * * * * * * *
85% * * * * * * * *       85% * * * * * * * *       85% * * * * * * * *
80% * * * * * * * *       80% * * * * * * * *       80% * * * * * * * *
75% * * * * * * * *       75% * * * * * * * *       75% * * * * * * * *
70% * * * * * * * *       70% * * * * * * * *       70% * * * * * * * *
 0% 6 1 1 * * * * *        0% 6 1 1 * * * * *        0% 6 * * * * * * *
   ---------------           ---------------           ---------------
    0 6 6 7 7 8 8 9           0 6 6 7 7 8 8 9           0 6 6 7 7 8 8 9
    0 0 5 0 5 0 5 0%          0 0 5 0 5 0 5 0%          0 0 5 0 5 0 5 0%

  MAZATEC.DATA (13)         MAZATEC.ADJUST (13)       MAZATEC.EXCLUDE (11)
90% * * * * * * * *       90% * * * * * * * *       90% * * * * * * * *
85% * * * * * * * *       85% * * * * * * * *       85% * * * * * * * *
80% * * * * * * * *       80% * * * * * * * *       80% * * * * * * * *
75% * * * * * * * *       75% * * * * * * * *       75% * * * * * * * *.
70% 8 8 8 8 8 9 * *       70% 8 8 8 8 8 9 * *       70% 8 8 8 8 8 9 * *
 0% 6 6 6 6 6 7 * *        0% 6 6 6 6 6 7 * *        0% 7 7 7 7 6 7 * *
   ---------------           ---------------           ---------------
    0 6 6 7 7 8 8 9           0 6 6 7 7 8 8 9           0 6 6 7 7 8 8 9
    0 0 5 0 5 0 5 0%          0 0 5 0 5 0 5 0%          0 0 5 0 5 0 5 0%

 POLYNESIA.DATA (69)       POLYNESIA.ADJUST (69)     POLYNESIA.EXCLUDE (59)
90% * * * * * * * *       90% * * * * * * * *       90% * * * * * * * *
85% * * * * * * * *       85% * * * * * * * *       85% * * * * * * * *
80% * * * * * * * *       80% * * * * * * * *       80% * * * * * * * *
75% - - - - * * * *       75% - - - - * * * *       75% * * * * * * * *
70% - - - - * * * *       70% - - - - * * * *       70% * * * * * * * *
 0% 6 5 4 3 5 * * *        0% 6 5 4 3 5 * * *        0% 7 6 5 4 5 * * *
   ---------------           ---------------           ---------------
    0 6 6 7 7 8 8 9           0 6 6 7 7 8 8 9           0 6 6 7 7 8 8 9
    0 0 5 0 5 0 5 0%          0 0 5 0 5 0 5 0%          0 0 5 0 5 0 5 0%

  SIOUAN.DATA (20)          SIOUAN.ADJUST (20)        SIOUAN.EXCLUDE (15)
90% * * * * * * * *       90% * * * * * * * *       90% * * * * * * * *
85% * * * * * * * *       85% * * * * * * * *       85% * * * * * * * *
80% - - - - - - - -       80% - - - - - - - -       80% * * * * * * * *
75% - - - - - - - 0       75% - - - - - - - 0       75% 5 5 5 5 5 5 5 5
70% - - - - - - - 0       70% - - - - - - - 0       70% 5 5 5 5 5 5 5 5
 0% 7 7 7 7 7 7 6 0        0% 7 7 7 7 7 7 6 0        0% 8 8 8 8 8 8 7 0
   ---------------           ---------------           ---------------
    0 6 6 7 7 8 8 9           0 6 6 7 7 8 8 9           0 6 6 7 7 8 8 9
    0 0 5 0 5 0 5 0%          0 0 5 0 5 0 5 0%          0 0 5 0 5 0 5 0%

   TRIQUE.DATA (11)          TRIQUE.ADJUST (11)         TRIQUE.EXCLUDE (7)
90% * * * * * * * *       90% * * * * * * * *       90% * * * * * * * *
85% * * * * * * * *       85% * * * * * * * *       85% * * * * * * * *
80% 6 6 6 6 6 5 * *       80% 6 6 6 6 6 5 * *       80% * * * * * * * *
75% 6 6 6 6 6 5 * *       75% 6 6 6 6 6 5 * *       75% * * * * * * * *
70% 7 7 7 7 7 5 * *       70% 7 7 7 7 7 5 * *       70% * * * * * * * *
 0% 6 6 6 6 6 5 * *        0% 6 6 6 6 6 5 * *        0% 8 8 8 8 8 8 * *
   ---------------           ---------------           ---------------
    0 6 6 7 7 8 8 9           0 6 6 7 7 8 8 9           0 6 6 7 7 8 8 9
    0 0 5 0 5 0 5 0%          0 0 5 0 5 0 5 0%          0 0,5 0 5 0 5 0%
```



Of the eleven sets of scores presented in Tables 2 and 3, most of the data that contribute to high correlations come in the area of low vocabulary similarity predicting low intelligibility. To the extent that adjustments based on hypotheses about how home town scores work and exclusions of large asymmetries make any difference at all, they also make it mostly in this same area.

Higher on the scale, where marginal intelligibility is involved, two sets of scores out of the eleven, those for Mazatec and Trique, show a strong overall correlation (taking the pocket of high correlations with adequate intelligibility in the Philippines as something with no parallel anywhere else). Yuman has a stronger correlation than Mazatec in the area of marginal intelligibility, but only when asymmetries are excluded.

It may be significant that Mazatec and Trique comparative linguistics is fairly well advanced. Gudschinsky, who studied Mazatecan, and Longacre, who studied Trique, had both been in vigorous debate with Morris Swadesh about the validity of lexical similarity measures in general. They and Paul Kirk, who did the Mazatec survey with Casad, commanded both a working knowledge and a comparativist's broad grasp of the dialects that went far beyond what might turn up by the luck of the draw on a survey word list; they knew about cognates that might never be noticed without the detailed comparative work they had already done.

Discussion

To the extent that the areas for which we already have data on vocabulary similarity and intelligibility represent dialect areas in the world in general, chances are 4.5 to 1 against any survey that attempts to assess intelligibility solely on the basis of vocabulary similarity being able to do so with any confidence, on the basis of 9 areas of low correlation against 2 of higher correlation.[9] One has to do the intelligibility testing anyway, not only because it is the best indicator we have of areas of high communication potential, but in a secondary sense in order to validate

whether the dialect area might indeed be one of the less likely ones in which vocabulary similarity bears a significant relation to comprehension.

This does not mean that we abolish the use of word lists in language surveys. It means instead that we no longer try to squeeze out of them information they are inherently incapable of giving. They do show up areas where intelligibility is unlikely, the ones where similarity is below 60%. Above that, counts based on them are helpful mainly to point up the need for intelligibility testing, but they are not a substitute for it.

Word lists should be used instead -- especially now that we have rapid methods for establishing consistencies in sound correspondence -- to give an initial picture of language groupings based on shared innovations in sound change, and to show the specific sound changes that result in those groupings. For such groupings, based on demonstrable genetic divergence, we can if we like quantify the conclusions reached by the comparative method, whether through phonostatistical indices of divergence,[10] groupings based on shared rules,[11] or computations of vocabulary similarity based on the retention of proven cognates under various conditions that encourage or discourage borrowing.

But even quantifications of full-fledged comparisons do not measure the other factors that are known to influence intelligibility. We have no comparable measures yet for calibrating morphological differences, syntactic differences,[12] or shifts of meaning, or for the social and geographic factors; nor if we had them would they necessarily combine meaningfully with a similarity index into a single composite figure that we could validate against measures of comprehension. For the present, however, we do have an effective strategy for arriving at decisions about language programs[13] that gives reasonable results unless we try to cut corners with it:

(1) Inspect word lists.
(2) If similarity is below 60%, assume separate programs.
(3) If similarity is 60% or better, test for intelligibility after screening subjects for possible bilingual learning of the other dialect.
(4) If intelligibility scores are uniform (standard deviation below 15%) and average scores are 85% and above, combined programs may be possible.
(5) If combined programs are linguistically possible, test social attitudes to make sure a combined program is feasible. The sampling requirements and the testing strategy for questionable cases are very different from the ones appropriate for testing for inherent intelligibility.
(6) If intelligibility scores are spread out (standard deviation 15% or greater), the problem becomes one of assessing what proportion of the population is at each level of bilingual proficiency. The sampling requirements and the testing strategy for determining this are very different from those appropriate for inherent intelligibility.

APPENDIX

Computer program in BASIC to calculate correlations.
Runs on the Sharp PC-5000, Kaypro 2000,
and other IBM PC compatibles.

```
1  'SIMINT - Correlate vocabulary similarity with intelligibility
2  'Joseph E. Grimes, 1986 October 13
3  '
```

```
10 LI=100 : LN=10 : L=60
20 DIM SD(LI), ID(LI)
30 DIM SL(LN), IL(LN), RROW(LN), NROW(LN)
40 NS=9 : NI=7
50 '
100 FOR I=1 TO NS : READ SL(I) : NEXT I 'Similarity thresholds
110  DATA 0,60,65,70,75,80,85,90,95
120 FOR I=1 TO NI : READ IL(I) : NEXT I 'Intelligibility thresholds
130  DATA 95,90,85,80,75,70,0
140 '
200 INPUT "File or device name for output";O$
210  OPEN O$ FOR OUTPUT AS #2
215  IF O$="LPT1" OR O$="lpt1" THEN PRINT#2,CHR$(27);"*1"
220 '
230 FOR Z=0 TO 1 STEP 0
235  IF O$<>"LPT1" AND O$<>"lpt1" THEN BEEP
240  INPUT "File name and extension [NNNNNNNN.XXX] for the data";F$
250  IF F$="" THEN CLOSE #2 : STOP
260   OPEN F$ FOR INPUT AS #1
270   '
300  FOR K=1 TO LI                'Read the data
310   INPUT #1, SD(K), ID(K)
320   IF SD(K)=0 AND ID(K)=0 THEN 350
330  NEXT K
340  K=K+1
350  K=K-1                        'K pairs read in
360  CLOSE #1
370  '
400  PRINT#2,TAB(4);F$;TAB(5+NS*5+3+4);F$ : L=L-1
410  PRINT#2,"Intelligibility>=I%"; TAB(5+NS*5+3); "Intelligibility>=I%" :
                        L=L-1
420  FOR I=1 TO NI
430   IN=IL(I)                    'By intelligibility thresholds
440   PRINT#2,USING"##% ";IN;
450   GOSUB 500                   'Correlate with similarity
460  NEXT I
470  GOSUB 900                    'Bottom of graph
480 NEXT Z                        'End of one data set
490 '
500 FOR J=1 TO NS                 'Correlate with similarity (450)
510  SIM=SL(J)                    'By similarity thresholds
520  GOSUB 600                    'Print the correlation
530  NROW(J)=N : RROW(J)=R        ' and save it for the graph
540 NEXT J
550 GOSUB 800                     'Print the graph
560 RETURN
570 '
600 N=0 : SX=0 : SY=0 : XY=0 : X2=0 : Y2=0 'Correlation (520)
610 FOR M=1 TO K                  'Go through the data
620  X=SD(M) : Y=ID(M)
630  IF X<SIM OR Y<IN THEN 650 ELSE N=N+1
640  SX=SX+X : SY=SY+Y : XY=XY+X*Y : X2=X2+X*X : Y2=Y2+Y*Y
650 NEXT M
660 '
```

```
700 IF N<5 THEN R=0 : PRINT#2,"    *"; : GOTO 780
710 A=SX/N : B=SY/N : C=XY/N : D=X2/N : E=Y2/N
720 YX=XY-(SX*SY) : XX=A*A-D
730 BB=YX/XX          'Slope of regression line
740 AA=B-BB*A         'Intercept of regression line
750 R=(XY-(SX*SY)/N) / SQR(ABS(X2-(SX*SX/N))*ABS(Y2-(SY*SY/N)))
                        'Correlation
760 IF R>1 OR R<-1 THEN PRINT#2,"   ??"; : GOTO 780
770 PRINT#2,USING" #.##";R;
780 RETURN
790 '
800 PRINT#2,USING"   ##%";IN;    'Print the graph (550)
810 FOR J=1 TO NS
820  IF NROW(J)<5 THEN PRINT#2," *"; : GOTO 850
830  IF RROW(J)<0 THEN PRINT#2," -"; : GOTO 850
840  IF RROW(J)>=0 THEN PRINT#2,USING" #";INT(RROW(J)*10);
850 NEXT J
860 PRINT#2,"" : L=L-1
870 RETURN
880 '
900 PRINT#2,TAB(10);                 'Bottom of graph (470)
910 FOR J=2 TO NS : PRINT#2,"-----"; : NEXT J
920 PRINT#2,TAB(5+NS*5+6);
930 FOR J=1 TO NS : PRINT#2,"--"; : NEXT J
940 PRINT#2,"" : L=L-1
950 '
1000 PRINT#2,TAB(5);
1010 FOR J=1 TO NS : PRINT#2,USING"  ##%";SL(J); : NEXT J
1020 PRINT#2,TAB(5+NS*5+6);
1030 FOR J=1 TO NS : PRINT#2,USING" #";INT(SL(J)/10); : NEXT J
1040 PRINT#2,"" : L=L-1
1050 '
1100 PRINT#2,TAB(5+NS*5+6);
1110 FOR J=1 TO NS : PRINT#2,USING" #";SL(J)-INT(SL(J)/10)*10; : NEXT J
1120 PRINT#2," %" : L=L-1
1130 PRINT#2,TAB(12);"Similarity>=S% for"; K; "pairs"; TAB(5+NS*5+6+3);
                        "Similarity>=S%"
1140 PRINT#2,TAB(5);"* N<5 too few to correlate   - R<0 reversed
                        0 to 9 R=.0 to .9"
1150 PRINT#2,"" : PRINT#2,"" : L=L-4
1160 IF L-NI-8<=0 THEN L=60 : PRINT#2,CHR$(12)
1170 RETURN
1180 END
```

NOTES

[1]Using word lists to study regularities in sound change is not a short cut comparison. It is, however, at least as time consuming and demanding as intelligibility testing, even with the aid of a computer.

[2]Most surveys fail to report either this range of individual variation or the number of speakers tested in each place. Both are needed in order to interpret correctly the average intelligibility, which is the only one of the three essential figures that usually appears The standard deviation is a useful measure of the range

of variation. It is easy to calculate and is needed for statistical reasoning. It is gotten by taking the amount by which each individual's score deviates from the average, squaring it to keep the deviations that are below the average from canceling out those above it, adding up the squares of the deviations, dividing the sum by the number of test subjects to get the average squared deviation, then taking the square root of that to put it all back onto the original scale. Standard deviations for inherent intelligibility are normally less than 15% (a ball park figure, not yet validated precisely), while a larger standard deviation is typical of bilingual situations.

[3]Casad, p. 177, gives a set of individual scores from the Mazatec survey. The scores for Huautla (Hu) break into three groups. Three people have scores in the fifties: I would guess they represent the real intelligibility. Four have scores 90 to 100: they could be the practiced bilinguals, though this kind of test cannot distinguish high fluency from only moderate proficiency. The other three are in between, possibly reflecting low proficiency in the Huautla dialect. Huautla is a market and cultural center whose speech is learned by people from the countryside. The mean intelligibility is 76%, but the standard deviation is 18%, enough of a scatter to raise suspicion.

[4]Three significant digits gives a spurious impression of the accuracy that can be attained from the form of the test usually given. Rounding the community averages to the nearest 5% would reflect the inherent precision of that type of test even better than rounding to 1%. Tests and sampling procedures could be devised that would be accurate to 1%, but they would be extremely costly, and the decisions about language programs made on that basis would not be noticeably different from tests accurate to 5%.

[5]There are, of course, many other dialect pairs in the Philippines for which neither similarity nor intelligibility figures have been compiled. The ones given represent dialects that were judged close enough to be worth testing. Most of the others fit into the low intelligibility, low similarity category mentioned in the first paragraph of the paper. They would fill the lower left quadrant of Figure 1.

[6]Positive correlations go with regression lines that rise from left to right like the one in Figure 1. If the line fell, indicating inverse correlation (the more of this, the less of that), the correlation would be negative, and perfect negative correlation would be expressed as -1.0.

[7]Only 1,448 correlations are actually given because the rows and columns that correspond to 95% were dropped from all the tables after they were calculated, since none contained enough data to give a valid correlation. Each correlation involved from 6 pairs of numbers for Biliau to 29 pairs for Polynesian.

[8]The problems introduced artificially by adjusting intelligibility scores that fall below 100% for the subjects' home towns have been greatly reduced by stipulating that the only questions considered admissible for a text be those on which the panel of home town speakers that Casad uses have a score of 100%. At the time he published his monograph it was still not clear that this would work. On pp. 61-62 he speaks of throwing out questions that half the speakers had difficulty with; the improvement has come from following through on this concept by throwing out all questions (out of a very large initial pool of possible ones) that the panel cannot answer well. In earlier testing, inability to answer questions in one's mother tongue had been thought to be due mainly to the unfamiliarity of the test situation, so the corrections proposed treated it as an error factor in learning that would diminish

with time and experience. For the argument I am making I have to assume that its noise effect is randomly distributed throughout these data, because we have no record of the order in which the test tapes used were presented to different subjects, and hence of how their responses might be weighted for learning behavior.

[9]The score is 8 to 3 if the Yuman data after exclusions are taken into account as well.

[10]See Grimes and Agard 1959 and Grimes 1964 for phonostatistical quantification.

[11]Suitable methods are described in Romesburg 1984, in sections on qualitative resemblance matrices.

[12]Andrew S. Noetzel and Stanley M. Selkow, and David Sankoff and Roger J. Cedergren, in two chapters in Sankoff and Kruskal 1983, lay the groundwork for tree-to-tree comparison measures.

[13]Barbara F. Grimes (1985) explains the rationale for this strategy.

REFERENCES

Casad, Eugene H. 1974. *Dialect Intelligibility Testing*, Summer Institute of Linguistics Publications in Linguistics and Related Fields, no. 38. Norman OK: Summer Institute of Linguistics of the University of Oklahoma.

Grimes, Barbara F., editor. 1984. *Ethnologue: Languages of the World*, tenth edition. Dallas TX: Wycliffe Bible Translators, Inc.

------. 1985. 'Comprehension and Language Attitudes in Relation to Language Choice for Literature and Education in Pre-Literate Societies'. *Journal of Multilingual and Multicultural Development* 6:2.165-181.

Grimes, Joseph E. 1964. 'Measures of Linguistic Divergence', Horace G. Lunt, ed., *Proceedings of the Ninth International Congress of Linguistics, Cambridge, Mass., 1962*. The Hague: Mouton. Pp. 44-50.

------ and Frederick B. Agard. 1959. 'Linguistic Divergence in Romance'. *Language* 35.598-604.

Romesburg, H. Charles. 1984. *Cluster Analysis for Researchers*. Belmont CA: Lifetime Learning Publications.

Sankoff, David, and Joseph B. Kruskal, eds. 1983. *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Reading MA: Addison-Wesley Publishing Company.

Simons, Gary F. 1979. *Language Variation and Limits to Communication*. Cornell University Department of Modern Languages and Linguistics. Reissued by the Summer Institute of Linguistics, Dallas TX.